

Task-Cloning Algorithms in a MapReduce Cluster with Competitive Performance Bounds

Huanle XU, Wing Cheong LAU

Department of Information Engineering, The Chinese University of Hong Kong
{xh112, wclau}@ie.cuhk.edu.hk

Abstract—Job scheduling for a MapReduce cluster has been an active research topic in recent years. However, measurement traces from real-world production environment show that the duration of tasks within a job vary widely. The overall elapsed time of a job, i.e. the so-called flowtime, is often dictated by one or few slowly-running tasks within a job, generally referred as the “stragglers”. The cause of stragglers include tasks running on partially/intermittently failing machines or the existence of some localized resource bottleneck(s) within a MapReduce cluster. To tackle this online job scheduling challenge, we adopt the task cloning approach and design the corresponding scheduling algorithms which aim at minimizing the weighted sum of job flowtimes in a MapReduce cluster based on the Shortest Remaining Processing Time scheduler (SRPT). To be more specific, we first design a 2-competitive offline algorithm when the variance of task-duration is negligible. We then extend this offline algorithm to yield the so-called SRPTMS+C algorithm for the online case and show that SRPTMS+C is $(1 + \epsilon) - \text{speed } o(\frac{1}{\epsilon^2}) - \text{competitive}$ in reducing the weighted sum of job flowtimes within a cluster. Both of the algorithms explicitly consider the precedence constraints between the two phases within the MapReduce framework. We also demonstrate via trace-driven simulations that SRPTMS+C can significantly reduce the weighted/unweighted sum of job flowtimes by cutting down the elapsed time of small jobs substantially. In particular, SRPTMS+C beats the Microsoft Mantri scheme by nearly 25% according to this metric.

Index Terms—MapReduce, job Scheduling, SRPT, cloning, weighted job flowtime, competitive bound

I. INTRODUCTION

MapReduce [8] and its open-source realization via Hadoop [1] have emerged as the defacto framework to support large-scale parallel/distributed processing and data analytics. Under the MapReduce framework, the overall computation of a job is decomposed into 2 separate phases, namely, the Map phase and the Reduce phase. Within each phase, many relatively small tasks are executed in parallel across a large number of machines within the MapReduce cluster. The MapReduce computational model also requires that the Reduce phase of a job cannot begin until all the tasks within its Map phase have been completed. A key feature of catalyzing the widespread adoption of MapReduce framework is the ability to transparently deal with the challenges of executing these tasks in a distributed setting. One of such fundamental challenges is the disproportionately long-running tasks, or the so called stragglers, which corresponding to tasks that are unfortunately assigned to machines suffering from partially/intermittently failures or localized resource bottleneck(s). Measurement traces from the

real-world production environment [4] indicate that stragglers lead to a large variation in completion times among tasks in the same job phase and delay job completion substantially.

The dominant technique to mitigate the straggler problem is via speculative execution: a strategy which preventively or reactively handle stragglers via automatically launching of extra copies of a task on alternative machines. In particular, there are two main classes of speculative execution strategies proposed in the literature, namely, the Cloning approach [2] and the Straggler-Detection-based one [1], [4], [7], [14], [28]. Under the Cloning approach, extra copies of a task are scheduled in parallel with the initial task and the one which finishes first is used for the subsequent computation. For the Straggler-Detection based approach, the progress of each task is monitored by the system and backup copies are launched when a straggler is detected. Unfortunately, most of these speculative execution schemes are based on simple heuristics and generally lack any performance guarantee.

To take a more systematic approach for the design of speculative execution strategies, our previous work (e.g., [24]–[26]) propose several optimization-based schemes: [26] proposes to make clones for each task of the arriving jobs by running a convex program which aims at minimizing the total job elapsed time, which is the time-span between the job arrival and its completion. This is commonly referred as the flowtime of a job in the scheduling literature. However, [24]–[26] still face two fundamental limitations. Firstly, the precedence constraints between the two phases in the MapReduce framework are ignored. Secondly, the complete distribution of task duration within each job needed to be known in advance when solving the optimization problem. Ideally, we want to take the precedence constraint into consideration and reduce the amount of information required for optimizing the speculative execution scheme.

With the above ideas in mind, in this paper, we explicitly model the precedence between the Map and Reduce phase and assume that only the first and second moments of task duration are known *a priori*. Similar to [26], we aim to minimize the weighted sum of job flowtime via task cloning. This objective yields offline and online versions of the scheduling problem which turns out to be more difficult than the NP-Hard scheduling problem presented in [29]. Our main results include the approximated algorithms which are motivated from the Shortest Remaining Processing Time scheduler (SRPT) in both offline and online setting. To be more specific, we

obtain a 2-competitive algorithm for the offline case when the variance of task-duration is negligible and a $(1 + \epsilon) - \text{speed}$ $o(\frac{1}{\epsilon^2}) - \text{competitive}$ algorithm for the online case where $0 < \epsilon < 1$. For the online version of the algorithm, we assume resource augmentation [15], which is necessary to circumvent lower bounds for the parallel scheduling on multiple machines. Under the resource augmentation analysis, the adversary is given m unit-speed machines and our algorithm is given M processors of speed s where $s > 1$. To summarize, this paper has made the following technical contributions:

- After reviewing the related work in Section II, we cast the dynamic scheduling problem as an stochastic optimization problem that focuses on finding a cloning scheme to minimize the weighted sum of job flowtimes (Section III).
- Motivated by the SRPT scheduler, we design a 2-competitive algorithm for the offline case when the variance of task duration is negligible. Moreover, we show that, with high probability, each job can complete within a time-span which is larger than the optimal algorithm by only a constant factor times the standard derivation of task duration (Section IV).
- Extended from the offline algorithm, we design the so-called SRPTMS+C algorithm for the online case. By adopting the method of potential function analysis, we prove that SRPTMS+C is $(1 + \epsilon) - \text{speed}$ $o(\frac{1}{\epsilon^2}) - \text{competitive}$ for the weighted sum of job flowtimes when $0 < \epsilon < 1$ (Section V).
- Before concluding our work in Section VII, we demonstrate via trace-driven simulations that SRPTMS+C can significantly reduce the weighted average of job flowtimes by cutting down the elapsed time of small jobs substantially. In particular, SRPTMS+C beats the Microsoft Mantri scheme by nearly 25% according to this metric (Section VI).

II. RELATED WORK

The straggler problem was first identified in the original MapReduce paper [8]. Since then, various solutions have been proposed to deal with it using the Straggler-Detection-based speculative execution strategy [4], [7], [14], [28]. These solutions mainly focus on promptly identifying stragglers and accurately predicting the performance of running tasks. One fundamental limitation is that detection may be too late for helping small jobs as it needs to wait for the collection of enough samples while monitoring the progress of tasks. To avoid the extra delay caused by the straggler detection, cloning approach was proposed in [2]. This approach relies on cloning very small job in a greedy manner to mitigate the straggler-effect and is based on simple heuristics. In contrast, we develop an optimization framework to make clones for each arriving job. Recently, [3] presents GRASS, which carefully adopts the Detection-based approach to trim stragglers for approximation jobs. GRASS also provides a unified solution for normal jobs. However, one limitation is that it only prioritizes the tasks within a job and it remains a problem

to prioritize different jobs (i.e., the scheduler is not optimized and unknown to the readers).

Prior research on job scheduling for a MapReduce system includes [5], [6], [17], [19], [22], [27], [29]: [5], [6], [27] derive performance bounds for minimizing the total completion time. [22] designs the *Coupling scheduler*, which mitigates the starvation problem caused by Reduce tasks in large jobs. [19], [27], [29] extend the SRPT scheduler to minimize the total job flowtime under different settings. However, all of these studies assume accurate knowledge of task durations and hence do not support speculative copies to be scheduled dynamically.

Finally, the SRPT scheduler has been studied extensively in traditional parallel scheduling literature. In particular, SRPT has proven to be $(1 + \epsilon) - \text{speed}$ $\frac{4}{\epsilon} - \text{competitive}$ for total flowtime on m identical machines under the single task case [11]. In this paper, we extend the SRPT scheduler to yield an online scheduler which can mitigate stragglers as well.

III. SYSTEM MODEL AND PROBLEM FORMULATION

Consider a MapReduce Cluster which consists of M machines. A machine could represent a processor, a core or a virtual machine. Assume a set of jobs $\mathcal{J} = \{J_1, J_2, \dots\}$ entering into the cluster over time. Job $J_i \in \mathcal{J}$ which arrives at the cluster at time a_i consists of m_i map tasks and r_i reduce tasks. Each job has a weight w_i which reflects its priority. Let $J_i^m = \{\delta_i^{m,1}, \delta_i^{m,2}, \dots, \delta_i^{m,m_i}\}$ and $J_i^r = \{\delta_i^{r,1}, \delta_i^{r,2}, \dots, \delta_i^{r,r_i}\}$ be the set of map and reduce tasks of J_i respectively. Each machine can only hold one map or reduce task at any time and all the machines are identical.

As described in Section I, the large variation in task completion times is caused by machine failures or localized resource bottleneck(s). Instead of modelling the variance of machine speed directly, we consider that the variation is caused by task workload and each machine processes all the tasks with the same speed. Such transformation does not violate the variation in task completion times and simply our analysis.

We assume time is slotted and a centralized scheduler collects the status of jobs within the cluster at the beginning of each time slot. If a machine runs a task at speed s , it will take $p(\cdot)/s$ time slots to complete the task where $p(\cdot)$ denotes the workload of this task. Without loss of generality, we assume that all the machines run at unit speed.

For ease of presentation, throughout the whole paper, we use $c \in \{m, r\}$ to capture the map- or reduce-related statements for all the tasks, i.e., when c is used, it is fixed to either m or r . The workload of task $\delta_i^{c,j} \in J_i^c$ is $p_i^{c,j}$ where $p_i^{c,j}$ is a random number for all i, j . Under the unit speed case, $p_i^{c,j}$ also denotes the processing time of task $\delta_i^{c,j}$ on a particular machine. We also assume the workload of all tasks in a job share the same mean E_i^c and standard deviation σ_i^c . The parameters E_i^c and σ_i^c are known in advance to the scheduler for all i .

Table I summarizes all the notations in this model.

A. Speedup via task cloning

In this model, we adopt the cloning approach to mitigate the negative impact of stragglers. Cloning helps to speed up the

Table I
THE NOTATIONS OF THE SCHEDULING PARAMETERS

Notations	Corresponding meaning
\mathcal{J}	The set of jobs arriving at the cluster
J_i^c	The set of map/reduce tasks of job J_i with $c = \{\text{map for map task; reduce for reduce task}\}$
a_i	Arrival time of job J_i
f_i	Time when job J_i completes its work
w_i	Weight of job J_i
$p_i^{c,j}$	Workload of the map/reduce task $\delta_i^{c,j}(\rho_i^j)$
E_i^c	The mean of the workload for map/reduce task in J_i
σ_i^c	The SD of the workload for map/reduce task in J_i
M	Total number of machines in the Cluster
$s_i^c(x)$	The speedup function of a map/reduce task in J_i
c_i^j	Time when task $\delta_i^{c,j}$ is scheduled.
$t_i^{c,j}$	The duration of task $\delta_i^{c,j}$.
$f_i^{c,j}$	Time when task $\delta_i^{c,j}$ completes.
$M(t)$	Number of machines running map tasks at time t .
$R(t)$	Number of machines running reduce tasks at time t .
$x_i^{c,j}$	Number of copies made for task $\delta_i^{c,j}$.

completion of a task via picking up the copy which finishes first of this task. To capture such speedup, we define a function, which is $s_i^c(x)$, for each phase of every single job where x is the number of copies made for a particular task. For example, it takes $p_i^{c,j}/s_i^c(2)$ time slots to complete task $\delta_i^{c,j}$ on average if two copies are made when scheduling $\delta_i^{c,j}$. Here, we assume that $s_i^c(x)$ satisfies the following two properties:

- $s_i^c(x)$ is a concave and strictly increasing function of x , $\forall i$.
- $s_i^c(1) = 1$ and $s_i^c(x) \leq x$ for all $x > 0$, $\forall i$.

These two properties are applicable to most distributions of the task duration observed in practice. For example, [4], [26] show that the task duration for a MapReduce cluster follows a heavy-tail distribution. Below, we illustrate the convexity of the speedup function when the task duration follows a Pareto heavy-tail distribution. In particular, if the duration $p_i^{c,j}$ of task $\delta_i^{c,j}$ follows the following Pareto distribution, we have:

$$Pr(p_i^{c,j} < t) = \begin{cases} 1 - (\frac{\mu}{t})^\alpha & \text{for } t \geq \mu \\ 0 & \text{otherwise} \end{cases}$$

when r copies are made for the task $\delta_i^{c,j}$, the average duration of $\delta_i^{c,j}$ is $\frac{\alpha r \mu}{\alpha r - 1}$. The derivation of this result is shown in [25]. As such, the speedup function is just $s_i^c(r) = \frac{r\alpha - 1}{r(\alpha - 1)}$ which is strictly concave and monotonic.

B. A stochastic program formulation for job scheduling

For any job, all the map tasks and reduce tasks can only be scheduled after the job arrival at the cluster and hence $m_i^j \geq a_i$. The Map phase ends when all the map tasks finish, i.e., $f_i^{m,j} = m_i^j + t_i^{m,j} \quad \forall i; 1 \leq j \leq m_i$. Due to the precedence constraints of the Map and Reduce phase, a reduce task can not begin its work if some map tasks within the same job do not finish. Thus, the reduce task $\delta_i^{r,j}$ can only start after the end of the Map phase (i.e., $\max\{\max_k\{f_i^{m,k}\}, r_i^j\}$). At any time slot, the total number of machines available for processing the tasks and their clones cannot exceed M , i.e., $M(t) + R(t) \leq M$.

Finally, a job completes when all the reduce tasks are finished, i.e., $f_i = \max_j\{f_i^{r,j}\} \quad \forall i; 1 \leq j \leq r_i$.

For this model, we aim to minimize the weighted sum of job flowtimes by carefully making cloning decisions and prioritizing different jobs. This formulation yields an optimization problem shown below:

$$\min \sum_i w_i \cdot \mathbb{E}[f_i - a_i] \quad (1a)$$

$$s.t. \quad m_i^j \geq a_i \quad \forall i; 1 \leq j \leq m_i \quad (1b)$$

$$r_i^j \geq a_i \quad \forall i; 1 \leq j \leq r_i \quad (1c)$$

$$\mathbb{E}[t_i^{m,j}] = E_i^m / s_i^m(x_i^{m,j}) \quad \forall i; 1 \leq j \leq m_i \quad (1d)$$

$$\mathbb{E}[t_i^{r,j}] = E_i^r / s_i^r(x_i^{r,j}) \quad \forall i; 1 \leq j \leq r_i \quad (1e)$$

$$f_i^{m,j} = m_i^j + t_i^{m,j} \quad \forall i; 1 \leq j \leq m_i \quad (1f)$$

$$f_i^{r,j} = \max\{\max_k\{f_i^{m,k}\}, r_i^j\} + t_i^{r,j} \quad \forall i; j \quad (1g)$$

$$\sum_{m_i^j \geq t; f_i^{m,j} < t} x_i^{m,j} = M(t) \quad \forall t \quad (1h)$$

$$\sum_{r_i^j \geq t; f_i^{r,j} < t} x_i^{r,j} = R(t) \quad \forall t \quad (1i)$$

$$M(t) + R(t) \leq M \quad \forall t \quad (1j)$$

$$f_i = \max_j\{f_i^{r,j}\} \quad \forall i; 1 \leq j \leq r_i \quad (1k)$$

Constraint (1d) and (1e) illustrate the speedup property for the map tasks and reduce tasks respectively. Constraint (1g) is due to the precedence constraint of the Map and Reduce phase.

Remark 1. When task cloning is not used and there is no variation in completion times among tasks in the same job phase, the scheduling problem in our model just reduces to the problem in [29]. However, the optimization problem presented in [29] has proven to be NP-Hard even for the offline case where all the jobs enter into the cluster at the same time. The stochastic optimization problem in Equation (1) therefore is NP-Hard and hence we resort to the use of approximation algorithms to tackle this problem.

IV. OFFLINE SCHEDULING: ALL THE JOBS ARRIVE AT THE CLUSTER AT THE SAME TIME

Before designing the online algorithm, in this section, we consider an offline case where all the jobs enter into the system at the same time, namely, $a_i = 0 \quad \forall i$. We assume that all the tasks cannot be launched simultaneously in the cluster (e.g., $\sum_{i=1} c_i > M$), otherwise, we can assign all the tasks to the machines in the cluster and the scheduling process just ends. Although this setting is simple, the offline algorithm presented below provides good insights for us to design an online algorithm in the following sections.

[3] builds a simple model to analyze the advantage of pure cloning. It concludes that cloning cannot help to reduce the job flowtime if $s_i^c(x) \leq x$ when the number of tasks to be scheduled is larger than M . Therefore, we do not clone extra copies for the tasks in this bulk arrival scenario.

A. Offline Algorithm Design

It is well known in scheduling literature that the SRPT scheduler is optimal for reducing overall flowtime on a single machine when there is only one task per job [20]. In each time slot, the SRPT scheduler always selects the job with the minimum total remaining workload to serve. We extend this SRPT scheduler to design the following offline scheduling algorithm:

In this algorithm, the scheduler first applies the SRPT scheduler for the scheduling problem in which there is only one single machine to determine the priority of each job. Let ϕ_i be the total effective workload of job J_i , which is determined by the following equation:

$$\phi_i = m_i \cdot (E_i^m + r\sigma_i^m) + r_i \cdot (E_i^r + r\sigma_i^r) \quad (2)$$

The standard deviation of task duration is incorporated into the workload via multiplying by a factor r and the priority of job J_i is then defined as w_i/ϕ_i . The rationale of including the standard derivation of task duration in the effective workload of task is that tasks with large variation in completion times can easily prolong the job completion and hence should be scheduled later. However, it still remains a problem to choose a good r and we tackle this problem in Section VI.

After computing the priority for each job, the jobs with higher priorities are always scheduled before those ones with lower priorities. Whenever a machine is available, the scheduler randomly chooses one unscheduled task from the pool of not-yet-finished jobs, or the set of alive jobs that keep the highest priority and assign it to this machine. Moreover, all the map tasks are scheduled before the reduce tasks in the same job. Since the Reduce phase can only begin after the Map phase finishes, a reduce task cannot make progress even after it has been scheduled as long as there are some unfinished map tasks within the same job.

Algorithm 1: Offline Scheduling algorithm for the bulk arrival

Input: The jobs associated with c_i , E_i^c and σ_i^c ;
Output: Allocated machines for all the map tasks and reduce tasks.

- 1 Sort the job set \mathcal{J} based on the decreasing order of w_i/ϕ_i ;
- 2 Initialize the job set $\Phi = \mathcal{J}$;
- 3 **if** A machine is available **then**
- 4 **for** each job J_i in Φ **do**
- 5 **if** J_i has unscheduled map task **then**
- 6 Choose one unscheduled map task at random and assign it to this machine;
- 7 **else**
- 8 Choose one unscheduled reduce task at random and assign it to this machine;
- 9 **if** J_i has no unscheduled task **then**
- 10 $\Phi = \Phi - \{J_i\}$;

Algorithm 1 presents the pseudo-code of the algorithm.

B. Deriving the upper bound for job flowtime

We proceed to analyze the performance of Algorithm 1. Define f_i^s as the accumulated workload of those jobs whose priority is larger than job J_i . In other words,

$$f_i^s = \sum_{j: w_j/\phi_j \geq w_i/\phi_i} \phi_j \quad (3)$$

We aim to show a generic bound on the flowtime of each job with a certain probability. To achieve this goal, we first prove the following lemma:

Lemma 1. *With probability at least $\frac{r^2-1}{r^2}$, the cluster is processing the jobs with priority at least w_i/ϕ_i during the interval $[0, f_i - E_i^r - r\sigma_i^r]$.*

Refer to Appendix A for the detailed poof of Lemma 1. Based on Lemma 1, we derive the following theorem which provides an upper bound for the flowtime of each job:

Theorem 1. *The flowtime of Job J_i is bounded by $E_i^r + r\sigma_i^r + f_i^s/M$ with a probability at least $1 + 1/r^4 - \frac{2}{r^2}$.*

Refer to Appendix B for the proof of Theorem 1.

Remark 2. *When the variance of the task workload is zero, the flowtime of each job is bounded by $E_i^r + f_i^s/M$ under Algorithm 1. Regardless of the type of scheduler, the flowtime of each job must be larger than E_i^r . On the other hand, the performance of the optimal scheduler is no better than the SRPT scheduler with one machine in terms of weighted sum of job flowtimes. Under the SRPT scheduler with one machine, the flowtime of each job is just f_i^s/M . Hence, we conclude that Algorithm 1 achieves a constant competitive ratio of two.*

However, if the variance is non-negligible, Algorithm 1 could not achieve a constant competitive ratio but still provides an upper bound for the flowtime of each individual job.

V. ONLINE SCHEDULING WITH CLONING FOR JOB ARRIVAL OVER TIME

In this section, we first present an approximated algorithm for the online scheduling case where all the jobs arrive at the cluster over time. After that, we provide an upper bound for the competitive ratio of the proposed algorithm.

A. Shortest Remaining Processing Time based Machine Sharing Principle

Extended from the offline algorithm presented in Section IV, we design a SRPT based machine sharing algorithm for this online case. The principle of machine sharing is motivated by the work in [9], [10], [12] where the machines are shared among the latest jobs arriving at the cluster (LAPS). A classic result in [9] shows that the LAPS algorithm is scalable for minimizing the total flowtime of jobs with sublinear speedup functions on multiple machines. However, in our algorithm, we share the machines among jobs with the smallest remaining workload. Other than that, we also make clones for the tasks according to machine availability.

We call this approximated algorithm the *Shortest Remaining Processing Time based Machine Sharing plus Cloning* (SRPTMS+C). At a high level, the SRPTMS+C algorithm works as follows: At the beginning of each time slot, the scheduler computes a priority for every alive job (i.e., not-yet-finished job). Let ϵ be a number such that $0 < \epsilon < 1$. Jobs with the highest priorities share the machines in proportion to their weights so that the weight of all running jobs is an ϵ fraction of the total weights of the alive jobs in the system. Observe that when ϵ is set to 1, the scheduler just reduces to the fair scheduler in Hadoop [1]. On the other hand, if ϵ is close to 0, the scheduler becomes the SRPT scheduler. By tuning the parameter ϵ , we could obtain a scheduler that best fits a cluster. More importantly, this ϵ fraction sharing principle yields a bounded competitive ratio as presented in the following sections.

Let $\psi^s(l)$ be the set of alive jobs at the beginning of time slot l . Denote by $m_i(l)$ and $r_i(l)$ the number of unscheduled map and reduce tasks of job J_i respectively. The remaining effective workload of job J_i can be characterized by:

$$U_i(l) = m_i(l) \cdot (E_i^m + r\sigma_i^m) + r_i(l) \cdot (E_i^r + r\sigma_i^r) \quad (4)$$

The scheduler computes $\frac{w_i}{U_i(l)}$ for each job in $\psi^s(l)$ and guarantees that the jobs with larger $\frac{w_i}{U_i(l)}$ have higher priorities to be scheduled. Let

$$W(l) = \sum_{i \in \psi^s(l)} w_i \quad (5)$$

and $\psi_i^s(l)$ be the set of jobs alive for SRPTMS+C at time slot l which have lower priorities (i.e., smaller $\frac{w_i}{U_i(l)}$) than J_i . J_i is also included in $\psi_i^s(l)$. Define $W_i(l) = \sum_{j \in \psi_i^s(l)} w_j$ and let

$$g_i(l) = \begin{cases} \frac{w_i \cdot M}{(\epsilon W(l))} & W_i(l) - w_i \geq (1 - \epsilon)W(l) \\ 0 & W_i(l) < (1 - \epsilon)W(l) \\ \frac{(W_i(l) - (1 - \epsilon)W_i(l)) \cdot M}{(\epsilon W(l))} & \text{otherwise} \end{cases}$$

Each job $J_i \in \psi^s(l)$ shares $g_i(l)$ machines within the cluster, including those ones that are still running the tasks of J_i , whose size is defined as $\sigma_i(l)$. Hence, the number of machines assigned to J_i in time slot l is $(g_i(l) - \sigma_i(l))$.

B. Task-Cloning Algorithm Design

When allocating the number of machines for each job (i.e., $(g_i(l) - \sigma_i(l))$), there may exist one case which violates the basic sharing principle in Section V-A, namely, the number of machines running the tasks of J_i (i.e., $\sigma_i(l)$) already exceeds $g_i(l)$ for some i . Under such situation, the scheduler reserves the work already completed for job J_i and just runs the tasks of J_i with their clones on $\sigma_i(l)$ machines. In other words, the scheduler does not allow preemption and lets J_i occupy these extra machines. Due to this non-preemptive mechanism, the exact number of machines shared by Job J_i may be larger than $g_i(l)$.

After the number of machines is allocated for each job, the scheduler needs to choose appropriate tasks of the alive jobs

Algorithm 2: SRPTMS+C Algorithm Design for On-line Scheduling

```

1 Update  $\psi^s(l)$ , the set of jobs which have unscheduled
  tasks at current time slot;
2 Update the number of available machines  $M(l)$ ;
3 Compute  $U_i(l)$  for each  $J_i \in \psi^s(l)$  and sort the jobs
  according to the decreasing order of  $\frac{w_i}{U_i(l)}$ ;
4 Compute  $W(l)$  based on Equation (5);
5 for the Job  $J_i \in \psi^s(l)$  do
6   Compute  $g_i(l)$ , the number of machines  $J_i$ 
   deserved according to the  $\epsilon$  fractional sharing
   policy;
7 for the Job  $J_i \in \psi^s(l)$  &&  $g_i(l) > 0$  do
8   Count the number of machines which still run the
   tasks of  $J_i$  including all the clones and denote it
   by  $\sigma_i(l)$ ;
9   Compute the number of newly available machines
   which is  $\xi_i(l) = g_i(l) - \sigma_i(l)$ ;
10  if  $\xi_i(l) \leq 0$  then
11    continue;
12  if  $\xi_i(l) < M(l)$  then
13    Assign  $\xi_i(l)$  extra machines to  $J_i$ ;
14    Call the task scheduling procedure for  $J_i$  with
     $\xi_i(l)$  machines with returning value  $\pi_i(l)$ ;
15     $M(l) -= \pi_i(l)$ ;
16  if  $\xi_i(l) \geq M(l)$  then
17    Assign  $M(l)$  extra machines to  $J_i$ ;
18    Call the task scheduling procedure for  $J_i$  with
     $M(l)$  machines with returning value  $\pi_i(l)$ ;
19     $M(l) -= \pi_i(l)$ ;
20 return;
```

Procedure Task Scheduling for Job J_i with x newly allocated machines

Input: The number of newly allocated machines x and the running status;

Output: Task scheduling decision for J_i and returning value $\pi_i(l)$.

```

1 Count  $m_i(l)$  and  $r_i(l)$ , the number of unscheduled
  map tasks and reduce tasks for  $J_i$  respectively;
2 if  $m_i(l) > 0$  &&  $m_i(l) \geq x$  then
3   run  $\lfloor x/m_i(l) \rfloor$  copies for each unscheduled task
  on available machines.
4   return  $x - \lfloor x/m_i(l) \rfloor * m_i(l)$ ;
5 else if  $m_i(l) > 0$  &&  $m_i(l) < x$  then
6   Choose  $x$  unscheduled map tasks uniformly at
  random and run one copy for each task on
  available machines;
7   return 0;
8 else
9   Repeat the same scheduling process for reduce
  tasks with  $x$  allocated machines.
```

for scheduling and make cloning decisions carefully. Following the precedence constraint of the Map and Reduce phase, the scheduler begins to schedule reduce tasks after all the map tasks completed. In addition, the clones are made for the tasks depending on whether the number of machines allocated to a particular job is larger than the number of unscheduled tasks. Take job J_i for example: When $g_i(l) - \sigma_i(l) > c_i(l)$, cloning will be made to fully utilize these machines allocated to J_i . To be more specific, the scheduler spawns the same number of clones for all the unscheduled tasks in $g_i(l)$. Otherwise, tasks with fewer clones are more likely to lag behind. Thus, each unscheduled task of J_i will be made $\lceil (g_i(l) - \sigma_i(l)) / c_i(l) \rceil$ copies. In contrast, when $g_i(l) - \sigma_i(l) \leq c_i(l)$, following the same argument of the offline scheduling algorithm, clones are not made in this case. Hence, the scheduler chooses some unscheduled tasks from $J_i(l)$ at random and launch it without cloning.

Algorithm 2 presents the pseudo-code of the algorithm.

C. Resource augmentation analysis

In this section, we use resource augmentation to analyze the performance of the SRPTMS+C algorithm. Before going to the details of the analysis, we first present the following definition which characterizes the performance of an approximated algorithm.

Definition 1. An approximated algorithm is s -speed c -competitive if the algorithm's objective is within a factor of c of the optimal solution's objective when the algorithm is given s resource augmentation [16].

Proposition 1. Consider any continuous and concave function $f : \mathbb{R}^+ \rightarrow \mathbb{R}^+$ with $f(0) \geq 0$. Then for any $b \geq a > 0$, we have $\frac{f(a)}{a} \geq \frac{f(b)}{b}$.

Proof: According to the definition of concave function, it holds that $f(\lambda x + (1 - \lambda)y) \geq \lambda f(x) + (1 - \lambda)f(y)$ for any $x, y \in \mathbb{R}^+$ and $\lambda \in [0, 1]$. Specially, consider $x = 0$, $y = b$ and $\lambda = 1 - \frac{a}{b}$. Then we have $f(a) \geq (1 - \frac{a}{b})f(0) + \frac{a}{b}f(b) \geq \frac{a}{b}f(b)$. Q.E.D. ■

Remark 3. Based on Proposition 1, we conclude that $f(\frac{1}{\epsilon} \cdot a) \geq \frac{1}{\epsilon} f(a)$, this can be proved by substituting $x = \frac{1}{\epsilon} \cdot a$ and $y = a$ into the inequality.

With the help of Proposition 1, we derive the following theorem which provides an upper bound for the competitive ratio of SRPTMS+C.

Theorem 2. The algorithm SRPTMS+C is $(1 + \epsilon) - \text{speed } o(\frac{1}{\epsilon^2}) - \text{competitive}$ for the expectation of weighted sum of job flowtimes when $0 < \epsilon < 1$.

The method of potential function analysis is widely adopted to derive performance bound with resource augmentation for online parallel scheduling algorithms in the literature (e.g., [9], [10], [13], [23]). The key step of this method is to define a proper potential function which combines the adversary and

Table II
GOOGLE TRACE DATA STATISTICS

Total number of Jobs	6064
Trace duration (s)	35032
Average number of tasks per job	26.31
Minimum task duration (s)	12.8
Maximum task duration (s)	22919.3
Average task duration (s)	1179.7

our algorithm. To be more specific, let $A(t)$ and $OPT(t)$ be the accumulated weighted sum of job flowtimes in the algorithm's and adversary's schedules, respectively. We define a potential function $\Phi(t)$ that satisfies the following properties which are extended from [13]:

- Boundary Condition: $\Phi(0) = \Phi(\infty) = 0$.
- Changes Condition when job arrives or completes: the value of the potential function decreases or remains the same when a job arrives or completes in our algorithm and the adversary.
- Dynamic Changes Condition: with ϵ resource augmentation, at any time when no job arrives or completes, $\mathbb{E}[\frac{dA(t)}{dt}] + \mathbb{E}[\frac{d\Phi(t)}{dt}] \leq \frac{c}{\epsilon^2} \mathbb{E}[\frac{dOPT(t)}{dt}]$.

By integrating over time, one can see that the existence of such a potential function is sufficient to yield a $(1 + \epsilon) - \text{speed } o(\frac{1}{\epsilon^2}) - \text{competitive}$ algorithm. Refer to Appendix C for the detailed proof.

VI. PERFORMANCE EVALUATION

In this section, we evaluate the performance of the SRPTMS+C algorithm via extensive simulations driven by Google cluster-usage traces [21]. The traces contain the information of job submission and completion time of Google services on a cluster of 12K servers. It also includes the number of tasks as well as the duration of each task. In addition, the priority for each job ranges from 0 to 11 and we just treat this priority as the job weight. From the traces, we extract the statistics of more than 6000 jobs during a 12-hour period. We already exclude those jobs which have specific constraints on machine attributes. The detailed job statistics are illustrated in Table II.

When running the simulations, we estimate the distribution for the workloads of all the tasks within each job phase. Once a cloning copy is made for a particular task, the workload for this clone is just drawn independently from the estimated distribution. We repeat the same simulation for each of the following evaluations ten times and take the average to obtain the final result.

A. Baseline Algorithms for comparison

We adopt the following two algorithms as the baselines to compare with the SRPTMS+C algorithm:

- **Microsoft Mantri's Speculative Execution Scheme:** The speculative execution scheme of Mantri is demonstrated to be the most effective one among all the straggler-detection based schemes [4]. Mantri estimates

¹ $\lceil x \rceil$ denotes the rounding of the real number x .

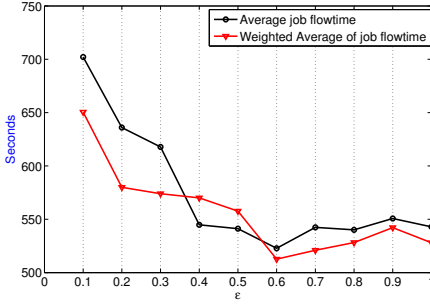


Figure 1. The weighted/unweighted average of job flowtimes for different ϵ under the SRPTMS+C algorithm when $r = 0$.

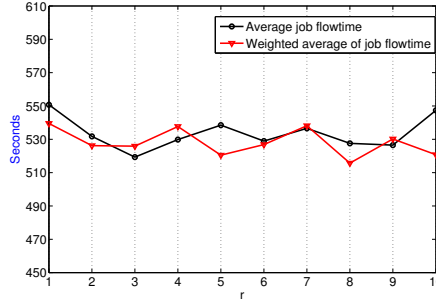


Figure 2. The weighted/unweighted average of job flowtimes for different r under the SRPTMS+C algorithm when $\epsilon = 0.6$.

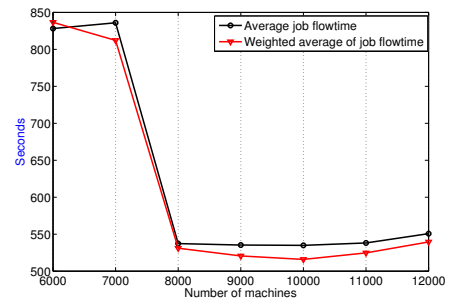


Figure 3. The weighted/unweighted average of job flowtimes under different number of machines for SRPTMS+C when $\epsilon = 0.6$ and $r = 3$.

the remaining time to finish, t_{rem} , for each task and predicts the required execution time of a relaunched copy of the task, t_{new} . Once a machine becomes available, the system makes a decision on whether to launch a backup copy based on the statistics of t_{rem} and t_{new} . Specifically, another copy is launched if the inequality $\mathbb{P}(t_{rem} > 2 * t_{new}) > \delta$ holds.

- **Smart Cloning Algorithm (SCA):** SCA is a cloning algorithm which is proposed in [26]. At the beginning of each time slot, SCA first runs a convex program to determine the number of copies assigned for each task and then launch all the copies simultaneously on available machines. SCA has been demonstrated to cut down the elapsed time of small jobs substantially.

Instead of comparing the weighted sum of job flowtimes directly, we take the weighted average for ease of presentation. Moreover, we also compare the unweighted average as well as the cumulative distribution function (i.e., CDF) of job flowtimes against different algorithms. The time scale of each slot is 1 second in our simulations.

B. The impact of ϵ , r and the number of machines in the cluster

In this subsection, we first evaluate the impact of ϵ and r on the average weighted/unweighted job flowtime under the SRPTMS+C algorithm in the cluster that contains 12K machines. Fig. 1 depicts the evaluation result under different ϵ when $r = 0$. Observe that when $\epsilon = 0.6$, which corresponding to the scheduler that schedules nearly half of the alive jobs with smaller effective workloads in each time slot, both of these two metrics attain the minimum.

To further evaluate the impact of r on the cluster performance, we set ϵ to 0.6 and evaluate the weighted/unweighted average of job flowtimes for different r under SRPTMS+C algorithm. It shows in Fig. 2 that the unweighted average of job flowtimes attain the minimum when $r = 3$ while the weighted average reaches its minimum under $r = 8$. In fact, both of these two metrics do not vary much between different r , the major reason is that the variation of task duration within each job phase for this particular job trace is small.

On the other hand, we scale out different number of machines in the cluster to show the impact on the job flowtime.

Observe from Fig. 3 that when the number of machines is around 8K, the performance is as equally well as it in the original cluster with 12K machines. There is enough resources to make clones for small jobs under the SRPTMS+C algorithm although the cluster only has 8K machines. The flowtime of small jobs from this trace therefore reduces substantially under SRPTMS+C.

C. Comparison against baseline algorithms

Based on the evaluation results above, we choose ϵ to be 0.6 and r to be 3 for the SRPTMS+C algorithm. We implement the three baseline algorithms as presented in their original papers in the cluster that contains 12K machines. The comparison results are illustrated in Fig. 4 and Fig. 5. Fig. 4 depicts the CDF of flowtime for the small jobs whose flowtime is between 0 and 300 seconds. It indicates that the SRPTMS+C algorithm obtains the best performance for those small jobs. In SRPTMS+C, more than 50% jobs complete within 100 seconds. In contrast, about 46% and 44% jobs complete within 100 seconds under SCA and Mantri respectively.

Fig. 5 depicts the CDF of flowtime for the big jobs whose flowtime are between 300 and 4000 seconds. One can see that the SRPTMS+C algorithm still achieves the best performance for these big jobs. For instance, about 90% jobs can complete within 1000 seconds under SRPTMS+C while only 88% and 86% jobs can complete within such time-span under SCA and Mantri respectively.

We illustrate the weighted/unweighted average of job flowtimes for this trace under different algorithms in Fig. 6. It shows that both of these two metrics under SRPTMS+C are reduced by nearly 25% comparing to Mantri baseline scheme. More importantly, the SRPTMS+C algorithm is much more efficient comparing to Mantri scheme in terms of implementation as the latter needs to monitor the progress of each running task which induces an extra system instrumentation.

VII. CONCLUSIONS

In this paper, we study the online scheduling problem in a MapReduce cluster and formulate a stochastic optimization program with the objective to minimize the weighted sum of job flowtimes. Following this model, we design the straggler-mitigation algorithms via task cloning, which are motivated

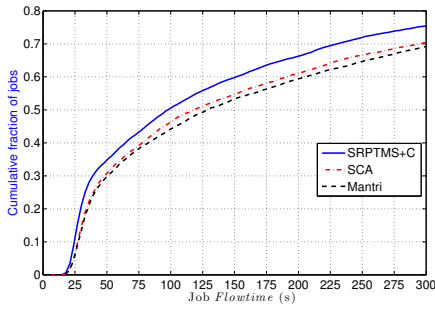


Figure 4. The cumulative fraction of the jobs within the flowtime ranging from 0 to 300 seconds under different algorithms.

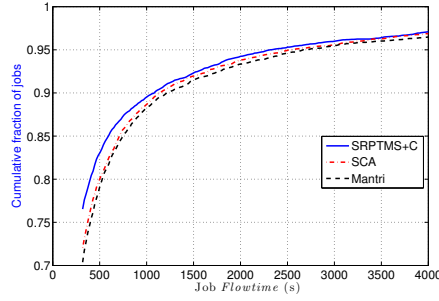


Figure 5. The cumulative fraction of the jobs within the flowtime ranging from 500 to 4000 seconds under different algorithms.

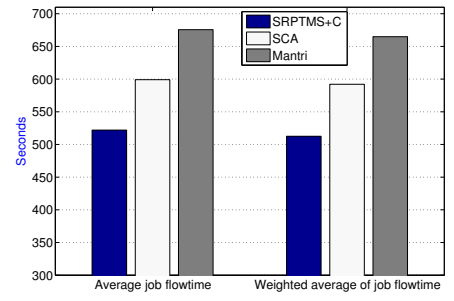


Figure 6. The weighted/unweighted average of job flowtimes under different algorithms within the cluster that has 12K machines.

from the SRPT scheduler in both offline and online cases. In the offline case, we show that, with high probability, each job can complete within a time-span which is larger than the optimal scheduling algorithm by only a constant factor times the standard derivation of task duration under our algorithm. When the variance of task duration is negligible, the offline algorithm achieves a competitive ratio of 2. On the other hand, we present the SRPTMS+C algorithm for the online case and provide an upper bound for the competitive ratio through the potential function analysis. Finally, we run several trace-driven simulations to evaluate the performance of the SRPTMS+C algorithm. It shows that SRPTMS+C cuts down the flowtime of small jobs substantially and reduces the wighted/unweighted sum of job flowtimes by nearly 25% comparing to Mantri baseline scheme.

REFERENCES

- [1] Apache. <http://hadoop.apache.org>, 2013.
- [2] G. Ananthanarayanan, A. Ghodsi, S. Shenker, and I. Stoica. Effective straggler mitigation: Attack of the clones. In *the 10th USENIX Symposium on Networked Systems Design and Implementation (NSDI)*, April 2013.
- [3] G. Ananthanarayanan, M. C.-C. Hung, X. Ren, I. Stoica, A. Wierman, and M. Yu. Grass: Trimming stragglers in approximation analytics. In *USENIX Symposium on Networked Systems Design and Implementation (NSDI)*, April 2014.
- [4] G. Ananthanarayanan, S. Kandula, A. Greenberg, I. Stoic, Y. Lu, B. Saha, and E. Harris. Reining in the outliers in MapReduce clusters using mantri. In *USENIX OSDI*, Vancouver, Canada, October 2010.
- [5] H. Chang, M. Kodialam, R. R. Kompella, T. V. Lakshman, M. Lee, and S. Mukherjee. Scheduling in MapReduce-like systems for fast completion time. In *Proceedings of IEEE Infocom*, March 2011.
- [6] F. Chen, M. Kodialam, and T. Lakshman. Joint scheduling of processing and shuffle phases in MapReduce systems. In *Proceedings of IEEE Infocom*, March 2012.
- [7] Q. Chen, C. Liu, and Z. Xiao. Improving MapReduce performance using smart speculative execution strategy. *IEEE Transactions on Computers*, PP(99), January 2013.
- [8] J. Dean and S. Ghemawat. MapReduce: Simplified data processing on large clusters. In *Proceedings of OSDI*, pages 137–150, December 2004.
- [9] J. Edmonds and K. Pruhs. Scalably scheduling processes with arbitrary speedup curves. In *Proceedings of SODA*, January 2009.
- [10] K. Fox, S. Im, and B. Moseley. Energy efficient scheduling of parallelizable jobs. In *Proceedings of SODA*, January 2013.
- [11] K. Fox and B. Moseley. Online scheduling on identical machines using srpt. In *ACM-SIAM Symposium on Discrete Algorithms (SODA)*, January 2011.
- [12] A. Gupta, S. Im, R. Krishnaswamy, B. Moseley, and K. Pruhs. Scheduling jobs with varying parallelizability to reduce variance. In *Proceedings of SPAA*, June 2010.
- [13] S. Im, B. Moseley, and K. P. an dEric Torng. Competitively scheduling tasks with intermediate parallelizability. In *Proceedings of SPAA*, June 2014.
- [14] M. Isard, M. Budiu, Y. Yu, A. Birrell, and D. Fetterly. Dryad: distributed data-parallel programs from sequential building blocks. In *Proceeding of the 2nd ACM SIGOPS/EuroSys European Conference on Computer Systems*, March 2007.
- [15] B. Kalyanasundaram and K. Pruhs. Speed is as powerful as clairvoyance. In *Proceedings of FOCS*, October 1995.
- [16] B. Kalyanasundaram and K. Pruhs. Speed is as powerful as clairvoyance. *Journal of the ACM*, 47:214–221, 1995.
- [17] M. Lin, L. Zhang, A. Wierman, and J. Tan. Joint optimization of overlapping phases in MapReduce. In *Proceedings of IFIP Performance*, September 2013.
- [18] M. Mitzenmacher and E. Upfal. *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press, January 2005.
- [19] B. Moseley, A. Dasgupta, R. Kumar, and T. Sarlos. On scheduling in Map-Reduce and flow-shops. In *Proceedings of SPAA*, pages 289–298, June 2011.
- [20] M. L. Pinedo. *Theory, Algorithms, and Systems*. Springer Publishing Company, 2008.
- [21] C. Reiss, J. Wilkes, and J. L. Hellerstein. Google cluster-usage traces. <http://code.google.com/p/googleclusterdata>, May 2011.
- [22] J. Tan, X. Meng, and L. Zhang. Delay tails in MapReduce scheduling. In *Proceedings of SIGMETRICS*, pages 5–16, London, United Kingdom, June 2012.
- [23] K. F. University and B. Moseley. Online scheduling on identical machines using SRPT. In *Proceedings of SODA*, January 2011.
- [24] H. Xu and W. C. Lau. Resource optimization for speculative execution in a MapReduce cluster. In *Proceedings of ICNP*, October 2013.
- [25] H. Xu and W. C. Lau. Speculative execution for a single job in a MapReduce-like system. In *International conference on Cloud Computing*, June 2014.
- [26] H. Xu and W. C. Lau. Optimization for speculative execution in a MapReduce-like cluster. To appear in Infocom, 2015.
- [27] Y. Yuan, D. Wang, and J. Liu. Joint scheduling of mapreduce jobs with servers: Performance bounds and experiments. In *Proceedings of IEEE Infocom*, April 2014.
- [28] M. Zaharia, A. Konwinski, A. D. Joseph, R. Katz, and I. Stoica. Improving Mapreduce performance in heterogeneous environments. In *Proceeding of the 8th USENIX conference on Operating systems design and implementation*, December 2008.
- [29] Y. Zheng, N. Shroff, and P. Sinha. A new analytical technique for designing provably efficient MapReduce schedulers. In *Proceedings of IEEE Infocom*, Turin, Italy, April 2013.

APPENDIX

A. Proof of Lemma 1

Proof: Consider the following case where the cluster is processing some jobs with priorities smaller than w_i/ϕ_i during the interval $[0, f_i - E_i^r - r\sigma_i^r]$. In this case, the job J_i must have already scheduled all the reduce tasks at time $f_i - E_i^r -$

$r\sigma_i^r$ according to the offline algorithm we present. Further, we consider the reduce task which is finished at last in job J_i and let $\delta_i^{r,j}$ denote it. According to the definition of f_i , $\delta_i^{r,j}$ finishes its work at time f_i . It implies that the workload of $\delta_i^{r,j}$ is at least $E_i^r + r\sigma_i^r$. Applying the Chebyshev Inequality [18] here, we get the following formula:

$$\Pr\{r_i^j \geq E_i^r + r\sigma_i^r\} \leq \Pr\{|r_i^j - E_i^r| \geq r\sigma_i^r\} \leq \frac{1}{r^2} \quad (6)$$

This completes the proof. \blacksquare

B. Proof of Theorem 1

Proof: Denote by X the work that the cluster has processed for the jobs with priority at least w_i/ϕ_i . Thus, the following two equations hold:

$$E[X] = \sum_{j:w_j/\phi_j \geq w_i/\phi_i} m_j \cdot E_j^m + r_j \cdot E_j^r. \quad (7)$$

$$\sigma^2[X] = \sum_{j:w_j/\phi_j \geq w_i/\phi_i} m_j \cdot (\sigma_j^m)^2 + r_j \cdot (\sigma_j^r)^2. \quad (8)$$

Applying the Chebyshev Inequality again, we conclude that the probability that X is no less than f_i^s is bounded by

$$\begin{aligned} & \Pr\{X \geq f_i^s\} \\ & \leq \Pr\left\{|X - E[X]| \geq r \sum_{j:w_j/\phi_j \geq w_i/\phi_i} (m_j \sigma_j^m + r_j \sigma_j^r)\right\} \\ & \leq \frac{\sum_{j:w_j/\phi_j \geq w_i/\phi_i} m_j \cdot (\sigma_j^m)^2 + r_j \cdot (\sigma_j^r)^2}{[r \sum_{j:w_j/\phi_j \geq w_i/\phi_i} (m_j \cdot \sigma_j^m + r_j \cdot \sigma_j^r)]^2} \\ & \leq \frac{1}{r^2} \end{aligned} \quad (9)$$

Applying Lemma 1 here, with probability at least $\frac{r^2-1}{r^2}$, the cluster is processing the work of X during the interval $[0, f_i - E_i^r - r\sigma_i^r]$. There are M machines processing the task with unit speed in total. Thus, with probability at least $(r^2-1)^2/r^4$, the following inequality holds.

$$M * (f_i - E_i^r - r\sigma_i^r) \leq f_i^s \quad (10)$$

The theorem immediately follows. Q.E.D. \blacksquare

C. Proof of Theorem 2

In the algorithm design of Section V-A, we assume that time is slotted. For convenience of analysis, here we consider a more general case where the time is continuous. In fact, we just make the length of a time slot small enough, as long as the duration of each task is the multiples of a slot length, our analysis doesn't violate the algorithm setting.

Proof: Let $y_i^j(t) = \max\{p_i^{Aj}(t) - p_i^{Oj}(t), 0\}$ where $p_i^{Oj}(t)$ and $p_i^{Aj}(t)$ denote the remaining workload to be processed for task δ_i^j in Job J_i at time t under the optimal scheduling policy and SRPTMS+C algorithm respectively. Let $\psi^o(t)$ and $\psi^s(t)$ be the jobs and tasks that are still alive (have

not completed yet) at time t in the optimal scheduling. Further denote by $\psi^s(t)$ the set of jobs that are alive in SRPTMS+C.

Denote by $t_i^{s,j}$ and $t_i^{f,j}$ the start time and completion time of task δ_i^j respectively. Based on the constraints (1d) and (1e), it follows that

$$\mathbb{E}[t_i^{f,j} - t_i^{s,j}] = \mathbb{E}[t_i^j] / s_i(x_i^j) \quad (11)$$

moreover, we have

$$\mathbb{E}\left[\int_{t_i^{f,j}}^{t_i^{s,j}} dp_i^{Aj}(t)\right] = \mathbb{E}[t_i^j] \quad (12)$$

Substituting Equation (11) into Equation (12) yields the following formula:

$$\mathbb{E}\left[\frac{dp_i^{Aj}(t)}{dt}\right] = -s_i(x_i^j) \quad (13)$$

The potential function for a single task is defined as follows:

$$\varphi_i^j(t) = \frac{w_i y_i^j(t)}{s_i(w_i M / \varepsilon W(t))} \quad (14)$$

Our overall potential function for all the jobs in the cluster is defined as

$$\Psi(t) = \frac{1}{\varepsilon^2} \sum_{i \in \psi^s(t)} \sum_{\delta_i^j \in J_i^c(t)} \varphi_i^j(t) \quad (15)$$

The Potential function is differentiable and it holds that $\Psi(0) = \Psi(\infty) = 0$ and

$$\mathbb{E}\left[\frac{d\Psi(t)}{dt}\right] = \frac{1}{\varepsilon^2} \sum_{i \in \psi^s(t)} \sum_{\delta_i^j \in J_i^c(t)} \mathbb{E}\left[\frac{d\varphi_i^j(t)}{dt}\right] \quad (16)$$

Let C_i^A be the completion time for Job J_i under SRPTMS+C algorithm. For any time $t \geq r_i$, let $A_i(t) = w_i(\min\{C_i^A, t\} - r_i)$ be the accumulated weighted flow time of Job J_i at time t , then we must have

$$\mathbb{E}\left[\frac{dA_i(t)}{dt}\right] = w_i \quad \text{for } r_i < t < C_i^A \quad (17)$$

$A_i(\infty)$ is just the flowtime of Job J_i in the cluster. Hence, the total flowtime of all the jobs in the cluster can be formulated as $A = \sum_i A_i(\infty)$. Further let $A_i = A_i(C_i^A)$ and $A(t) = \sum_i A_i(t)$. Let $\text{OPT}_i(t)$, OPT_i , $\text{OPT}(t)$ and OPT be defined similarly for the optimal scheduling policy.

Similar to the potential-function based analysis [9], [10], [13], [23], our goal is to bound the continuous and discrete increases to $\Psi(t)$ by a function of OPT .

We now focus on the changes made to $\Psi(t)$. It's obvious that the job arrivals make no change to this metric. In addition, the completion of jobs in the optimal schedule has no effect on the potential function value. The completion of jobs in SRPTMS+C causes the corresponding term being removed from $\Psi(t)$, however, it only decreases the potential and we just omit it as our goal is to obtain the upper bound for the changes made to $\Psi(t)$. As a result, we only need to analyze the continuous change to $\Psi(t)$.

- Changes in $\Psi(t)$ due to the optimal scheduling policy which is define as $\Delta^O(t)$:

Let a_i^{Oj} be the number of machines assigned to task δ_i^j of Job J_i in the optimal scheduling policy. Based on Equation (13) and the definition of potential function, the contribution made by the optimal scheduling to $\frac{d}{dt}\mathbb{E}[\varphi_i^j(t)]$ is bounded by the following formula:

$$\varepsilon^2 \Delta_i^{\text{Oj}} = \frac{w_i s_i(a_i^{\text{Oj}})}{s_i(w_i M / \varepsilon W(t))} \quad (18)$$

There are two categories for a_i^{Oj} which are $a_i^{\text{Oj}} \leq w_i M / \varepsilon W(t)$ and $a_i^{\text{Oj}} > w_i M / \varepsilon W(t)$. For the former case, applying the monotonic property of s_i function for all i , we have $\frac{w_i s_i(a_i^{\text{Oj}})}{s_i(w_i M / \varepsilon W(t))} \leq w_i$. For the latter case, applying Proposition 1, we get

$$\frac{w_i s_i(a_i^{\text{Oj}})}{s_i(w_i M / \varepsilon W(t))} \leq w_i \frac{a_i^{\text{Oj}}}{w_i M / \varepsilon W(t)} = \varepsilon W(t) \frac{a_i^{\text{Oj}}}{M} \quad (19)$$

Combining the two cases, it follows that

$$\varepsilon^2 \Delta_i^{\text{Oj}}(t) \leq \max \left\{ w_i, \varepsilon W(t) \frac{a_i^{\text{Oj}}}{M} \right\} \quad (20)$$

$$\leq w_i + \varepsilon W(t) \frac{a_i^{\text{Oj}}}{M} \quad (21)$$

which indicates

$$\begin{aligned} \Delta^O(t) &= \sum_{i \in \psi^o(t) \cap \psi^s(t)} \sum_{\delta_i^j \in J_i^c(t)} \Delta_i^{\text{Oj}}(t) \quad (22) \\ &\leq \frac{1}{\varepsilon^2} \sum_{i \in \psi^o(t) \cap \psi^s(t)} \sum_{\delta_i^j \in J_i^c(t)} w_i \\ &\quad + \frac{W(t)}{M\varepsilon} \sum_{i \in \psi^s(t)} \sum_{\delta_i^j \in J_i^c(t)} a_i^{\text{Oj}} \quad (23) \end{aligned}$$

For the first term of Equation (23), it follows that $\sum_{\delta_i^j \in J_i^c(t)} w_i \leq C w_i$ where C is the maximum number of copies made for each task in the optimal scheduling algorithm. Hence,

$$\begin{aligned} \frac{1}{\varepsilon^2} \sum_{i \in \psi^o(t) \cap \psi^s(t)} \sum_{\delta_i^j \in J_i^c(t)} w_i &\leq \frac{C}{\varepsilon^2} \sum_{i \in \psi^o(t) \cap \psi^s(t)} w_i \\ &\leq \frac{C}{\varepsilon^2} \sum_{i \in \psi^o(t)} w_i \\ &= \frac{C}{\varepsilon^2} \cdot \mathbb{E} \left[\frac{d\text{OPT}(t)}{dt} \right] \quad (24) \end{aligned}$$

For the second term in Equation (23), we have

$$\sum_{i \in \psi^s(t)} \sum_{\delta_i^j \in J_i^c(t)} a_i^{\text{Oj}} \leq M \quad (25)$$

and

$$W(t) = \sum_{i \in \psi^s(t)} w_i = \sum_{i \in \psi^s(t)} \mathbb{E} \left[\frac{dA_i(t)}{dt} \right] = \mathbb{E} \left[\frac{dA(t)}{dt} \right] \quad (26)$$

Substitute Equation (24), (25) and (26) into Equation (23), it yields that

$$\Delta^O(t) \leq \frac{C}{\varepsilon^2} \cdot \mathbb{E} \left[\frac{d\text{OPT}(t)}{dt} \right] + \frac{1}{\varepsilon} \mathbb{E} \left[\frac{dA(t)}{dt} \right] \quad (27)$$

We proceed to analyze the changes to $\Psi(t)$ made by our SRPTMS+C scheduling.

- Changes in $\Psi(t)$ due to the SRPTMS+C scheduling policy which is defined as $\Delta^S(t)$:

For each task that is alive in SRPTMS+C at time t , if it completes the work in the optimal scheduling policy, then $y_i^j(t)$ is positive. Hence, $y_i^j(t)$ decreases for all tasks $\delta_i^j \notin \psi_s^o(t)$ that SRPTMS+C processes at time t .

We run our algorithm at speed of $1 + \varepsilon$. Let a_i^{Sj} be the number of machines assigned to task δ_i^j in SRPTMS+C at time t . According to our scheduling policy, we have $\sum_j a_i^{\text{Aj}} \leq g_i(t)$ for all $J_i \in \psi^s(t)$. It follows that

$$\begin{aligned} \Delta^S(t) &\leq -\frac{1+\varepsilon}{\varepsilon^2} \sum_{i \in \psi^s(t)} \sum_{\delta_i^j \in J_i^c(t) \cap \delta_i^j \notin \psi_s^o(t)} \frac{w_i s_i(a_i^{\text{Sj}})}{s_i(w_i M / \varepsilon W(t))} \\ &\leq -\left(\frac{1+\varepsilon}{\varepsilon} \right) W(t) \sum_{i \in \psi^s(t)} \sum_{\delta_i^j \in J_i^c(t) \cap \delta_i^j \notin \psi_s^o(t)} \frac{a_i^{\text{Sj}}}{M} \\ &= -\frac{(1+\varepsilon)W(t)}{\varepsilon} \left(\sum_{\substack{i \in \psi^s(t) \\ \delta_i^j \in J_i^c(t)}} \frac{a_i^{\text{Sj}}}{M} - \sum_{\delta_i^j \in \psi_s^o(t)} \frac{a_i^{\text{Sj}}}{M} \right) \\ &= -\left(\frac{1+\varepsilon}{\varepsilon} \right) W(t) \sum_{i \in \psi^s(t)} \frac{g_i(t)}{M} \\ &\quad + \left(\frac{1+\varepsilon}{\varepsilon} \right) W(t) \sum_{\delta_i^j \in \psi_s^o(t)} \frac{a_i^{\text{Sj}}}{M} \quad (28) \end{aligned}$$

The second inequality in the above follows Proposition 1 and the fact that $a_i^{\text{Sj}} \leq g_i(t) = \frac{w_i M}{\varepsilon W(t)}$. To bound Inequality (28), we need to bound the second term as follows:

$$\sum_{\delta_i^j \in \psi_s^o(t)} a_i^{\text{Sj}} \leq \sum_{i \in \psi^o(t)} \frac{w_i M}{\varepsilon W(t)} \quad (29)$$

$$\leq \frac{M}{\varepsilon W(t)} \mathbb{E} \left[\frac{d\text{OPT}(t)}{dt} \right] \quad (30)$$

In addition, we have

$$\sum_{i \in \psi^s(t)} g_i(t) = M \quad (31)$$

Substitute Inequality (29), (30) and Equation (31) into Inequality (28), it holds that

$$\begin{aligned} \Delta^S(t) &\leq -\left(\frac{1+\varepsilon}{\varepsilon} \right) W(t) + \left(\frac{1+\varepsilon}{\varepsilon^2} \right) \mathbb{E} \left[\frac{d\text{OPT}(t)}{dt} \right] \\ &= -\left(\frac{1+\varepsilon}{\varepsilon} \right) \mathbb{E} \left[\frac{dA(t)}{dt} \right] \\ &\quad + \left(\frac{1+\varepsilon}{\varepsilon^2} \right) \mathbb{E} \left[\frac{d\text{OPT}(t)}{dt} \right] \quad (32) \end{aligned}$$

We proceed to complete the final analysis based on the results derived above. Due to the fact that $\int_0^\infty \mathbb{E} \left[\frac{d\Psi(t)}{dt} \right] dt = \mathbb{E}[\Psi(\infty)] - \mathbb{E}[\Psi(0)] = 0$, we have

$$\begin{aligned}
\mathbb{E}[A] &= \int_0^\infty \mathbb{E} \left[\frac{dA(t)}{dt} \right] dt + \int_0^\infty \mathbb{E} \left[\frac{d\Psi(t)}{dt} \right] dt \\
&\leq \int_0^\infty \mathbb{E} \left[\frac{dA(t)}{dt} \right] dt + \int_0^\infty (\Delta^O(t) + \Delta^S(t)) dt \\
&\leq \int_0^\infty \mathbb{E} \left[\frac{dA(t)}{dt} \right] dt + \int_0^\infty \left(-\mathbb{E} \left[\frac{dA(t)}{dt} \right] \right) dt \\
&\quad + \int_0^\infty \left(\frac{C+1+\varepsilon}{\varepsilon^2} \mathbb{E} \left[\frac{d\text{OPT}(t)}{dt} \right] \right) dt \\
&= \left(\frac{C+1+\varepsilon}{\varepsilon^2} \right) \mathbb{E}[\text{OPT}]
\end{aligned} \tag{33}$$

This completes the proof. ■